SPEECH RECOGNITION DEVICE AND SPEECH RECOGNITION METHOD

Patent Number: JP2001255886
Publication date: 2001-09-21
Inventor(s): SUZUKI SATORU; OONO TAKEO; KIMURA TATSUYA
Applicant(s): MATSUSHITA ELECTRIC IND CO LTD
Requested Patent: JP2001255886
Application Number: JP20000064919 20000309
Priority Number(s):
IPC Classification: G10L15/10; G10L11/06; G10L15/02; G10L15/20; G10L21/02
EC Classification:
Equivalents:

Abstract

PROBLEM TO BE SOLVED: To provide a speech recognition device by which a user does not utter specific contents is not and which uses a speaker normalization processing capable of speedingly performing normalization into individual features of a speaker without requiring an on-line 'non-teacher'. SOLUTION: Featured values such as LPC cepstrum coefficients are extracted with the speech digitalized by A/D conversion as input signals (S10). Then, frequency axis conversion is conducted for the featured value such as the LPC cepstrum in order to normalize the effect caused by the individuality of the length of the vocal track of the uttering person (S30). Then, a matching is conducted between the featured values of the inputted speech that is frequency axis converted and the acoustic model featured values beforehand learned from plural speakers (S50). After that, the inputted utterings are made as teacher's signals based on the recognition result computed in the S50 and optimum conversion coefficients are obtained (S60). Then, smoothing is conducted for the conversion coefficients to absorb dispersion caused by the speakers and the phonemes and new updated frequency axis conversion coefficients are obtained (S70).

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－255886
（P2001－255886A）
(43)公開日　平成13年９月21日(2001.9.21)

(51)Int.Cl.⁷	識別記号	ＦＩ		ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ１０Ｌ 15/10		Ｇ１０Ｌ 101:16		５Ｄ０１５
11/06		9/16	３０１Ｂ	９Ａ００１
15/02		3/00	５１５Ｄ	
15/20		3/02	３０１Ａ	
21/02				

審査請求　未請求　請求項の数９　ＯＬ　（全 13 頁）　最終頁に続く

(21)出願番号　特願2000－64919(P2000－64919)

(22)出願日　平成12年３月９日(2000.3.9)

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　鈴木　哲
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

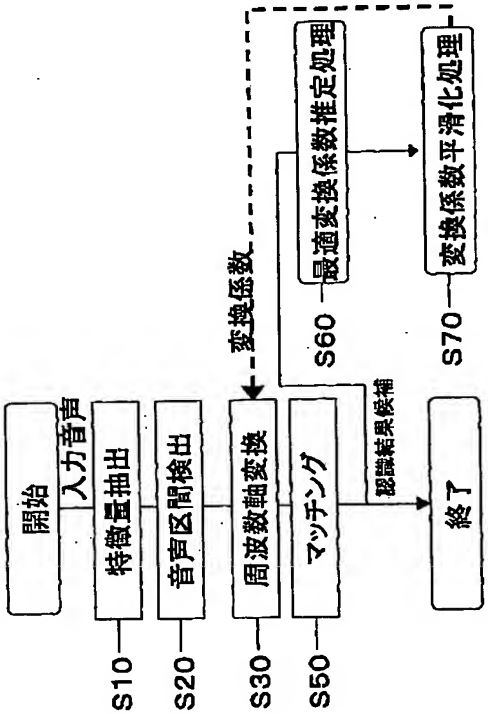
(72)発明者　大野　剛男
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

(74)代理人　100097445
弁理士　岩橋　文雄　（外２名）

最終頁に続く

(54)【発明の名称】　音声認識方法および音声認識装置

(57)【要約】
【課題】　利用者に対して一定内容の発声を促す必要性が無い、オンライン「教師なし」で話者の個人性特徴にすばやく正規化できる話者正規化処理を用いる音声認識装置を提供する。
【解決手段】　Ａ／Ｄ変換を行ってデジタル化された音声を入力信号として、ＬＰＣケプストラム係数等の特徴量を抽出し（Ｓ１０）、発声者の声道長の個人性に起因する影響を正規化するために、ＬＰＣケプストラム等の特徴量に周波数軸の変換を施し（Ｓ３０）、周波数軸変換を施された入力音声の特徴量と予め複数話者から学習した音響モデル特徴量とのマッチングを行なう（Ｓ５０）。その後、Ｓ５０において算出された認識結果をもとに入力発声を教師信号として最適な変換係数を求め（Ｓ６０）、話者や音韻によるばらつきを吸収するため変換係数平滑化を行い、新たな周波数軸変換係数を更新する（Ｓ７０）。



【特許請求の範囲】

【請求項１】　入力音声の特徴量を抽出する特徴量抽出ステップと、前記入力音声の特徴量の周波数軸を少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を用いて変換する周波数軸変換ステップと、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチングステップと、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定ステップと、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化ステップとを有することを特徴とする音声認識方法。

【請求項２】　変換係数平滑化ステップは、少なくとも１つの周波数軸変換係数から構成される最新の周波数軸変換係数列と、保持された過去に求められた少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列とを比較することによって話者の交代を検知することを特徴とする請求項１記載の音声認識方法。

【請求項３】　入力音声の特徴量を抽出する特徴量抽出ステップと、前記入力音声から無声音／有声音区間を弁別検出する音声区間検出ステップと、前記入力音声の特徴量の周波数軸を前記無声音／有声音区間情報に応じて周波数軸変換係数列を用いて変換する周波数軸変換ステップと、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチングステップと、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定ステップと、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化ステップとを有することを特徴とする音声認識方法。

【請求項４】　最適変換係数推定ステップは、前記認識結果候補を求めた際に使用した少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を基に複数の周波数軸変換係数列候補を設定し、それぞれの周波数軸変換係数毎に、前記マッチングステップの認識結果候補のうち少なくとも１つから表現される音素系列に対して、周波数軸変換を施して得られた入力音声特徴量と予め複数の話者から学習した音響モデル特徴量とマッチングにより尤度を求め、求めた尤度のうちで最大尤度を与える少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を選択することを特徴とする請求項１から３のいずれかに記載の音声認識方法。

【請求項５】　変換係数平滑化ステップは、未知話者に最適な周波数軸変換係数を平滑化する際に、保持された過去に求められた周波数軸変換係数の平均値と最新の周波数軸変換係数との距離を求め、所定の距離を満たす場合にのみ最新の係数として採用し、新たな周波数軸変換係数として出力することを特徴とする請求項１から４のいずれかに記載の音声認識方法。

【請求項６】　変換係数平滑化ステップは、未知話者に最適な周波数軸変換係数を平滑化する際に、保持された過去に求められた周波数軸変換係数の平均値と最新の周波数軸変換係数との距離を求め、所定の距離を満たさない場合には、過去の周波数軸変換係数を任意の初期値を与えることによって初期化を行うことを特徴とする請求項２記載の音声認識方法。

【請求項７】　入力音声の特徴量を抽出する特徴量抽出手段と、前記入力音声の特徴量の周波数軸を周波数軸変換係数列を用いて変換する周波数軸変換手段と、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチング手段と、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定手段と、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化手段とを有することを特徴とする音声認識装置。

【請求項８】　変換係数平滑化手段は、少なくとも１つの周波数軸変換係数から構成される最新の周波数軸変換係数列と、保持された過去に求められた少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列とを比較することによって話者の交代を検知することを特徴とする請求項７記載の音声認識装置。

【請求項９】　入力音声の特徴量を抽出する特徴量抽出手段と、前記入力音声から無声音／有声音区間を弁別検出する音声区間検出手段と、前記入力音声の特徴量の周波数軸を前記無声音／有声音区間情報に応じて周波数軸変換係数列を用いて変換する周波数軸変換手段と、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチング手段と、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定手段と、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化手段とを有することを特徴とする音声認識装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、不特定話者の音声を音声認識させる分野等に利用される音声認識方法および音声認識装置に関する。

【０００２】

【従来の技術】人の音声の音響的特徴は、主に音を発声させる喉つまり音源と、その音が反響しながら伝播する声道およびその形状とで構成される発声器官によって決定される。つまり話者の音響的特徴の違いは、これら話者の発声器官の特徴が主に起因して生じていると考える事ができる。

【０００３】そのため、特に不特定話者を対象とした音声認識装置において、音響モデルで表現される話者集団の発声器官の特徴と比較して特異な特徴を持つ話者の認識率が低下することがあると指摘されていた。

【０００４】そこで、この発声器官の特徴つまり個人性を要因とした認識率の低下を防ぎ、かつ高い認識率を保持することを目的として、話者適応化手法あるいは話者正規化手法が提案されてきた。

【０００５】従来提案されてきた話者適応化、話者正規化手法としては、音響モデルパラメータを既知の音響モデルパラメータを用いて再評価することにより音響モデル自身を話者にあわせて更新あるいは選択する話者適応方法と、個々の話者の特徴空間を変換して音響モデル学習話者から表現される特徴空間にマッピングする話者正規化方法との２つに大別できる。

【０００６】前者は、たとえば電子通信情報学会SP92―16(1992年)に紹介されたベクトル場平滑化法のように、適応化音声の量が増すに従い、話者適応システム性能は話者依存での学習時の性能に近づくという特徴を持つため格段の認識性能の向上が期待できるものの、適応の効果が現れるのに十分な学習音声量を獲得するまでに時間を要するという欠点がある。

【０００７】後者は、たとえば論文「Frequency Warpingによる話者の正規化、松本、脇田、日本音響学会音声研究会資料Ｓ79－25,1979-7」においては、周波数正規化スペクトルによって声道長正規化に効果があると主張しており、さらに、論文「ＬＰＣ距離尺度における周波数正規化に関する検討、小林、松本、日本音響学会講演論文集1-1-5、昭和58年10月」においては、ＬＰＣスペクトルの周波数軸を伸縮する方法として一次の全域通過フィルターを用いた方法を提案している。

【０００８】この後者の方法は、変換係数の変更により話者の個人性を正規化できるという特徴を持つことから、オンラインの話者適応化・正規化方法としては前者に比べて、必要とする音声量がより少ない点で実用上有効であると考えられる。さらに、話者正規化手法として、特開平６－２１４５９６号公報において、声帯音源特性に関する音声スペクトル傾斜の変動と、声道特性（例えば声道長）に関する音声スペクトルの周波数軸方向の伸縮という個人性を同時に正規化する方法が提案されている。

【０００９】以下、この従来例の音声認識装置について図９を用いて説明する。

【００１０】図９の音声認識装置は、入力された音声信号の周波数特性を補正する周波数特性補正部１０と、入力音声信号のケプストラム係数を入力音声特徴量として抽出する特徴量抽出部２０と、入力音声信号に対し周波数軸の変換を施す周波数軸変換部３０と、入力された音声信号の区間を検出する音声区間検出部４０と、標準音声信号の特徴量が標準音声特徴量として予め記憶されている標準音声記憶部５０と、入力音声信号に対し周波数特性補正部１０、特徴量抽出部２０、周波数軸変換部３０により得られた入力音声特徴量と標準音声記憶部５０に記憶されている標準音声特徴量との照合（マッチング）を行なうマッチング部６０とから構成されている。

【００１１】ところで、この音声認識装置では、不特定話者の音声をも良好に認識させることを目的として、実際の音声認識処理とその音声認識処理の開始に先立って話者適応学習処理がなされる。この２種類の処理を１つの装置で行なわせるため、図９の装置には、この装置の動作、機能を話者適応フェーズと音声認識フェーズとのいずれかに切換えるためのフェーズ選択部９０がさらに設けられている。

【００１２】また、これと関連させて、標準音声記憶部５０には、話者適応処理用の標準音声特徴量と音声認識用の標準音声特徴量とが記憶されている。また、周波数特性補正部１０には、話者適応学習用に、互いに異なる複数の周波数特性補正係数が予め用意され、また、周波数軸変換部３０には、話者適応学習用に、互いに異なる複数の周波数軸変換係数が用意されている。

【００１３】次に、話者適応フェーズについて説明する。話者適応フェーズにおいては、未知話者に既知の発声内容を発声させるようになっており、周波数特性補正部１０、周波数軸変換部３０では、この音声信号に対して、各々、複数の周波数特性補正係数、複数の周波数軸変換係数を順次に変えて処理を行ない、マッチング部６０は、それぞれの場合について、周波数特性補正部１０、特徴量抽出部２０、周波数軸変換部３０により得られた入力音声特徴量を標準音声記憶部５０に記憶されている話者適応処理用の標準音声特徴量とマッチングして、各入力音声特徴量と標準音声特徴量との尤度を求め、そのうち最大尤度を与える周波数特性補正係数と周波数軸変換係数とを選択し決定するようになっている。

【００１４】次に、音声認識フェーズについて説明する。音声認識フェーズにおいては、未知話者（実際には、話者適応フェーズで入力を行なった話者）の未知の発声内容の音声信号に対して、周波数特性補正部１０、周波数軸変換部３０では、上記話者適応フェーズにおい

て選択、決定された周波数特性補正係数と周波数軸変換係数とに基づいて処理を行ない、マッチング部６０は、このようにして周波数特性補正部１０、特徴量抽出部２０、周波数軸変換部３０により得られた入力音声特徴量を標準音声記憶部５０に記憶されている音声認識用の標準音声特徴量とマッチングして、最大尤度を与える標準音声特徴量に対応した語を認識結果候補として出力するようになっている。

【００１５】

【発明が解決しようとする課題】この方法は、上記のように話者に発声内容を指定して発声を行わせることにより話者の個人性特徴を正規化させる最適な変換係数を推定する話者適応フェーズと、話者適応フェーズにおいて推定された変換係数を用いて未知内容の発声を認識する音声認識フェーズからなり、２つのフェーズを切り替えて使用するように構成されている。

【００１６】しかしながら、「教師あり条件」での話者正規化手法であるこの従来法は、事前に未知話者に対し発声語彙を指定し学習データとして収録する必要があるため利用者への負担増を招いている。そこで、この負担を取り除くために、話者に事前に発声を要求しない「教師なし」条件で、かつ即効性のある方法での話者正規化方法の確立が必要である。

【００１７】本発明では、発声者の発声器官の特徴つまり個人性を正規化することにより、ここでは声道長差に起因する影響を除去するために周波数軸変換を用いた「教師なし」条件での話者正規化を行う手法を述べる。具体的には、未知話者による内容未知の発声を用いて、周波数軸変換係数の精度よい推定方法を確立することで、教師なし条件の話者正規化方法を実現するものである。以下手法を実現するために生じる課題について述べる。

【００１８】従来法などでは、声道長の違いは音声スペクトルが伸縮する形で現れることに着目して、入力音声スペクトルの周波数軸を変換し、音響モデルなどの標準となる音声スペクトルとの差を吸収する周波数軸変換を用いた話者正規化は効果があることを提示している。このとき、周波数軸変換の際に与える周波数軸変換係数が、声道長の違いに相当するともいえる。これを声道長正規化と呼ぶ。

【００１９】一方、音声は、声道の形を変えることによって音韻の特徴を作り出されているため、同一話者であっても、発声される音韻によって声道長は異なっており、声道長推定値も一定の範囲で変化することが知られている「音声認識における個人差の学習法について、古井、日本音響学会音声研究会資料Ｓ７５－２５、1975-11」。つまり、発声される音韻によって、異なる話者ではもちろんのこと同一話者内でも変動するものと考えられる。

【００２０】したがって、声道長差に起因する音声スペクトルへの影響を除去するために、入力音声の周波数軸変換を行うにあたり、最適な周波数軸変換係数も、音韻によって変動していると考えられる。

【００２１】そのため、最適な周波数軸変換係数を推定し次回の発声に利用できるように「教師なし」条件での話者正規化を考えた場合、今回発声された音韻と次回発声される未知の音韻の違いによって、今回推定された最適な周波数軸変換係数が必ずしも次回の発声には最適とはいえず、このことが未知の発声内容に対応して正規化を行うことを難しくする要因となっている。したがって、個人性の特徴の一つである声道長正規化を行うにあたって、未知話者による発声内容が未知の入力音声を教師信号として用いて オンライン「教師なし」話者正規化を行うためには、周波数軸変換係数の推定精度をより高める推定方法の確立が必要となる。

【００２２】また、この従来法では、音声スペクトルに対して周波数軸を伸縮することによって、声道特性に関する個人性正規化を行うにあたって、この際入力音声区間全体に一律の変換係数を用いている。そのため、声道特徴に無関係な無声音の区間に対しても周波数軸変換を行うと、特徴量としての性質を失う原因になりかねず、認識結果に悪影響を及ぼすことも考えられる。そこで、入力音声区間全体に一律の変換係数を用いるのではなく、有声音区間に限って周波数軸変換を行うことにより、精度よく周波数軸変換係数の推定を行うことができると考えられる。

【００２３】本発明の目的は、上記の問題点を解決し、利用者に対して 予め発声を要求せず、発声ごとに話者の音声から個人性を精度よく推定することにより、未知話者による発声内容が未知の入力音声にすばやく話者正規化できる話者正規化方法を備えた音声認識装置を提供することである。

【００２４】

【発明が解決するための手段】本発明による音声認識方法および音声認識装置は、声道長差に起因するスペクトルの伸縮の影響を除去するため、入力音声のスペクトルに対して周波数軸変換を行なうことによる話者正規化方式を用いる。未知話者による発声内容が未知の入力音声を教師信号として、最尤推定により最適な周波数軸変換係数を決定した上で、音韻の違いによる周波数軸変換係数のばらつきを吸収するために、過去の周波数軸最適変換係数との平滑化を行うものである。さらに推定した周波数軸変換係数と過去の周波数軸最適変換係数とを比較することによって、話者が交代した場合とみなして平滑化を行うこともできる。

【００２５】同一話者内での声道長の変動には限界があるため、未知話者による発声内容が未知の入力音声を教師信号として推定された周波数軸変換係数は、ばらつきはあっても一定の範囲内に収束することが期待できる。しかしこのとき、次のような問題が考えられる。（1）

周波数軸変換係数を変化させた場合に、高い尤度をかせぐ周波数軸変換係数の範囲と、マッチング処理によって出力される認識結果候補が発声内容と合致するつまり認識正解する周波数軸変換係数の範囲とは必ずしも一致しない。（２）発声内容によって高い尤度をかせぐ周波数軸変換係数の範囲の分布が異なる。（３）マッチング処理によって出力される認識結果候補が発声内容と異なっている場合つまり誤認識した場合には、誤った内容に対して推定を行うことになってしまうため、マッチング処理によって出力される認識結果候補が発声内容として正しい場合つまり認識正解した場合に比べて、推定される周波数軸変換係数が異なる値になることがありうる。

【００２６】そこで、過去の発声より求めた周波数軸変換係数との平均して平滑化することにより、推定された周波数軸変換係数のばらつきを吸収して、現在の話者への最適周波数軸変換係数が求められるものと考える。

【００２７】さらに、話者が交代した場合には、前回の発声から推定した最適周波数軸変換係数と今回の発声から推定した最適周波数軸変換係数との差が大きくなることが考えられる。このことを利用して話者が交代した場合には、平滑化処理を初期化するなどのこれに対処を行うことが可能である。

【００２８】また、最適な周波数軸変換係数推定時には、音声区間検出手段から出力された無声音／有声音区間情報に同期して、周波数軸変換を行う。このため周波数軸変換係数推定にとって有効な音声区間あるいは音韻にのみ周波数軸変換を行うことから、精度よく周波数軸変換係数を推定できる。

【００２９】以上より、発声内容によらず、事前の発声を必要としないオンライン「教師なし」話者正規化方法を実現することが可能な高性能な音声認識装置を提供することができる。

【００３０】

【発明の実施の形態】本発明の請求項１に記載の発明は、入力音声の特徴量を抽出する特徴量抽出ステップと、前記入力音声の特徴量の周波数軸を少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を用いて変換する周波数軸変換ステップと、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチングステップと、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定ステップと、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化ステップとを有するもので、発声者の音声特徴量から周波数軸上に現れる個人性を吸収することにより、認識率の向上させる作用を有する。

【００３１】請求項２に記載の発明は、変換係数平滑化ステップは、少なくとも１つの周波数軸変換係数から構成される最新の周波数軸変換係数列と、保持された過去に求められた少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列とを比較することによって話者の交代を検知することを特徴とするもので、話者交代を検出した際には交代前話者の推定周波数変換係数の影響を受けないように周波数軸変換係数を初期化するなどによって、交代後話者への最適な周波数軸変換係数を新たに求め、話者間の周波数軸変換係数の差異による、認識率の低下を防ぐ作用を有する。

【００３２】請求項３に記載の発明は、入力音声の特徴量を抽出する特徴量抽出ステップと、前記入力音声から無声音／有声音区間を弁別検出する音声区間検出ステップと、前記入力音声の特徴量の周波数軸を前記無声音／有声音区間情報に応じて周波数軸変換係数列を用いて変換する周波数軸変換ステップと、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチングステップと、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定ステップと、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化ステップとを有することを特徴とするもので、発声内容によって変動する推定変換係数のばらつきを抑えることにより、周波数軸変換を用いた話者正規化のより高い効果を与える作用を有する。

【００３３】請求項４に記載の発明は、最適変換係数推定ステップは、前記認識結果候補を求めた際に使用した少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を基に複数の周波数軸変換係数列候補を設定し、それぞれの周波数軸変換係数毎に、前記マッチングステップの認識結果候補のうち少なくとも１つから表現される音素系列に対して、周波数軸変換を施して得られた入力音声特徴量と予め複数の話者から学習した音響モデル特徴量とマッチングにより尤度を求め、求めた尤度のうちで最大尤度を与える少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列を選択することを特徴とするもので、事前に発声を行い話者の個人性を学習する適応フェイズなどを設ずに、認識時の発声そのものから学習を行う「教師なし」話者正規化を実現する作用を有する。

【００３４】請求項５に記載の発明は、変換係数平滑化ステップは、未知話者に最適な周波数軸変換係数を平滑化する際に、保持された過去に求められた周波数軸変換係数の平均値と最新の周波数軸変換係数との距離を求め、所定の距離を満たす場合にのみ最新の係数として採

用し、新たな周波数軸変換係数として出力することを特徴とするもので、発声内容によって変動する推定変換係数のばらつきを抑えることにより、周波数軸変換を用いた話者正規化のより高い効果を与える作用を有する。

【００３５】請求項６に記載の発明は、変換係数平滑化ステップは、未知話者に最適な周波数軸変換係数を平滑化する際に、保持された過去に求められた周波数軸変換係数の平均値と最新の周波数軸変換係数との距離を求め、所定の距離を満たさない場合には、過去の周波数軸変換係数を任意の初期値を与えることによって初期化を行うことを特徴とするもので、発声内容によって変動する推定変換係数のばらつきを抑えることにより、周波数軸変換を用いた話者正規化のより高い効果を与える作用を有する。

【００３６】請求項７に記載の発明は、入力音声の特徴量を抽出する特徴量抽出手段と、前記入力音声の特徴量の周波数軸を周波数軸変換係数列を用いて変換する周波数軸変換手段と、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチング手段と、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して、少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定手段と、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化手段とを有することを特徴とするもので、話者の音声特徴量から周波数軸上に現れる個人性を吸収する話者正規化によって、認識率の向上させる作用を有する。

【００３７】請求項８に記載の発明は、変換係数平滑化手段は、少なくとも１つの周波数軸変換係数から構成される最新の周波数軸変換係数列と、保持された過去に求められた少なくとも１つの周波数軸変換係数から構成される周波数軸変換係数列とを比較することによって話者の交代を検知することを特徴とするもので、話者交代を検出した際には交代前話者の推定周波数変換係数の影響を受けないように周波数軸変換係数を初期化するなどによって、交代後話者への最適な周波数軸変換係数を新たに求め、話者間の周波数軸変換係数の差異による、認識率の低下を防ぐ作用を有する。

【００３８】請求項９に記載の発明は、入力音声の特徴量を抽出する特徴量抽出手段と、前記入力音声から無声音／有声音区間を弁別検出する音声区間検出手段と、前記入力音声の特徴量の周波数軸を前記無声音／有声音区間情報に応じて周波数軸変換係数列を用いて変換する周波数軸変換手段と、前記周波数軸変換を施した入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とをマッチングし、認識結果候補を出力するマッチング手段と、前記認識結果候補のうち少なくとも１つから表現される音素系列に対して、少なくとも１つの周波数軸変換係数から構成される最適な周波数軸変換係数列を推定する最適変換係数推定手段と、前記求められた最適な周波数軸変換係数列と保持された過去に求められた周波数軸変換係数列とを平滑化し、新たな周波数軸変換係数列を更新・保持する変換係数平滑化手段とを有することを特徴とするもので、発声内容によって変動する推定変換係数のばらつきを抑えることにより、周波数軸変換を用いた話者正規化のより高い効果を与える作用を有する。

【００３９】以下、本発明の実施の形態について図を用いて説明する。

【００４０】（実施の形態１）図１は、本発明の実施の形態１における音声認識装置のブロック図である。図１において、１は入力音声に対してＡ／Ｄ変換処理などを行う音声取り込み手段、２は音声の音響的特徴をモデル化した音響モデル、３は単語系列における単語間の関係をモデル化した言語モデル、４はデータやプログラム装置に入力する入力手段、５はデータやプログラムを記録するメモリ、６はプログラムにしたがってデータを処理したり装置全体を制御するＣＰＵ、７は認識結果候補を出力する出力手段である。

【００４１】図２は、本発明の音声認識装置の処理手順を示すフローチャートであり、この図を用いて音声認識装置の処理手順を説明する。

【００４２】入力音声の特徴量抽出が行われるＳ１０では、マイクロフォン等から取りこまれた音声にＡ／Ｄ変換を行ってデジタル化された音声を入力信号として、一定フレーム周期毎にＬＰＣメルケプストラム係数を出力する。具体的には、文献「音声認識、今井著、共立出版、１９９５年１１月２５日」などに示されているこの方法を利用して、プリエンファシス：$1-z^{-1}$、窓周期：２０ｍｓ、フレーム周期：１０ｍｓ、ＬＰＣ分析次数：１０次、ケプストラム分析次数：１０次としてＬＰＣメルケプストラム係数を出力する。入力音声はここでは、8kHzサンプリングされるものとする。

【００４３】音声区間検出処理が行われるＳ２０では、無声音／有声音区間などの検出を行なう。たとえば、入力音声信号のフレームパワーを求め、入力開始数フレームでの平均値を求めておき、その平均値に比べ、フレームパワーが2倍になったフレーム区間を有声音区間とする方法などを用いることにより、有声音区間とすることもできる。あるいは、入力音声に1500Hz-3400Hz通過高域フィルタを掛けて、同様にして高域音声入力が有ったことを示す高域区間を求めておき、これを利用することにより、有声音区間、無声音区間の判定を大まかな判定をすることができる。

【００４４】周波数軸変換処理が行われるＳ３０では、声道長の個人差に起因するスペクトルを伸縮することによって声道長正規化を行う周波数軸変換を施す。具体的

には、入力音声のスペクトルを表現しているＬＰＣメルケプストラム係数に対して、例えば（数１）で表わされる１次の全域透過フィルタＨ(ｚ）を作用させて、周波数軸の変換を行なう。この周波数軸変換の手法は、論文「LPC距離尺度における周波数正規化に関する検討、小林 松本 熊田、1983、日本音響学会音声研究会資料S83-47,1983 Dec. 22」に示されているものである。

【００４５】

【数１】

$$H(z) = (z^{-1} - \alpha) / (1 - \alpha z^{-1}) \qquad (-1 < \alpha < 1)$$

【００４６】また、実際に（数1）を用いて周波数軸を変換によって、スペクトル伸縮による周波数軸変換後のＬＰＣメルケプストラム係数の算出方法としては、たとえば、論文「Discrete representation of signals, Oppenheim and Johnson, Proc.IEEE,60,pp681-691,June1972」で示されている手法を用いる。

【００４７】なお、本実施例では、ＬＰＣメルケプストラムの次数を例えば１０次としたメル周波数変換の処理も同時に行なう。メル尺度を最も良く近似する周波数軸変換係数αの値は、サンプリング周波数８ｋHzの場合、α_b=０．３１５付近とされており、この値をαの値を基準として $\alpha=\alpha_b+0.05$、$\alpha=\alpha_b-0.05$などと指定してスペクトルを伸縮させる。ここで指定する周波数軸変換係数の値は、変換係数平滑化処理が行われるＳ７０から算出された値を用いる。

【００４８】このとき、周波数軸変換係数の推定にとって有効な区間あるいは音韻のみ周波数軸変換を行うことによって、精度よく変換係数を推定させるため、音声区間検出処理から出力された無声音／有声音区間情報に同期して周波数軸変換を行う。たとえば、音声区間検出処理の行われるＳ２０より得た有声音区間のフレームにのみ、周波数軸変換を実施する。

【００４９】マッチング処理が行われるＳ５０では、Ｓ３０において周波数軸変換された入力音声特徴量と予め複数の話者から学習した音響モデル特徴量とを、言語モデルとして表現される単語辞書等を用いてマッチングを行い、少なくとも１つの認識結果候補を出力する。なお、このマッチングは、例えば、端点フリーＤＰ（ダイナミック・プログラミング）マッチング法によりなされる。

【００５０】最適変換係数推定処理が行われるＳ６０では、マッチングにより出力される認識結果候補のうち少なくとも１つから表現される音素系列に対して、複数の周波数軸変換係数候補を設けて、それぞれの周波数軸変換係数毎に、周波数軸変換された入力音声の特徴量と予め複数の話者から学習した音響モデル特徴量とのマッチングを行い尤度を求めて、最尤推定により最大尤度を与える周波数軸変換係数を最適な周波数軸変換係数α_nとして決定し、平滑バッファに登録する。

【００５１】この時、最適変換係数推定処理が行われるＳ６０で推定される最適な周波数軸変換係数は発声内容によってばらつきが生じたり、不正解の音素系列に対して推定を行った場合には、必ずしも最適でない場合もありうることが問題となる。

【００５２】そこで、変換係数平滑化処理が行われるＳ７０では、推定した最適な周波数軸変換係数のばらつきを吸収するために、最適変換係数推定処理Ｓ６０で求められた周波数軸変換係数α_nと過去の周波数軸変換係数を記憶した平滑化バッファより読み出された、たとえば過去10回の周波数軸変換係数の平均により平滑化された周波数軸変換係数を算出し、新たな周波数軸変換係数α_0として更新・記憶する。

【００５３】次に、最適変換係数推定処理について、図３（ａ）（ｂ）を用いて詳細に説明する。図３（ａ）で示されるように、変換係数候補の値として、認識結果候補を求めた際用いた周波数軸変換係数α_0に対して、α_0-0.05、α_0、α_0+0.05の3点を設定し（Ｓ１０１）、それぞれ周波数軸変換を施した入力音声特徴量と認識結果候補第一位r(1)の単語の音素系列で表現される音響モデル特徴量系列とから尤度を求める（Ｓ１０２）。たとえば、周波数軸変換係数候補を、$x0=\alpha_0-0.05$、$x1=\alpha_0$、$x2=\alpha_0+0.05$ を設定し、それぞれ得られる尤度をy0, y1, y2とする。

【００５４】

【数２】

```
z1= (y1 - y0)/(x1-x0)
z2= (y2 - y0)/(x2-x0)
ak= (z2-z1)/(x2-x1)
bk= z1 - (x0+x1)*ak
```

【００５５】周波数変換係数およびその尤度からなる3点を用いて二次曲線に近似すると、その二次近似曲線がピークをとる周波数変換係数αxは、次のようにあらわされる（Ｓ１０３）。

【００５６】

【数３】

$$\alpha x = -bk / 2 / ak$$

【００５７】さらに、前記3点同様に周波数軸変換係数αxに対する尤度を求め（Ｓ１０４）、4点の中から最大尤度を与える周波数軸変換係数α_nを採用する（Ｓ１０５）ものである。

【００５８】次に、尤度距離の計算処理を図３（ｂ）の処理フローチャートを用いて説明する。

【００５９】変換係数候補の値として設定された周波数軸変換係数に対して、入力音声の特徴量に周波数軸変換を施す（Ｓ１０６）。前記の周波数軸変換を施した入力音声の特徴量と、認識結果候補第１位r(1)の単語の音素

系列から表現される音響モデル特徴量の系列とから尤度を求める（Ｓ１０７）。

【００６０】なお、これまで最適変換係数推定の説明において、認識結果候補のうち第一位の候補r(1)のみを用いたが、これを認識結果候補r(n)をｎ位まで利用して、最大尤度を与える周波数軸変換係数を採用することもできる。

【００６１】また、上記の特徴量を周波数軸変換する際に、音声区間検出処理が行われるＳ２０で算出される無声音／有声音などの音声区間情報に同期して、複数の周波数軸変換係数を用いることもできる。たとえば、無声音の区間にのみ周波数軸変換を適応したり、逆に有声音の区間にのみ周波数軸変換を適応したり、音声パワーの同じ区間ごとに異なる周波数軸変換係数を用いることもできる。

【００６２】なお、変換係数平滑化処理において、最適変換係数推定処理Ｓ６０で算出された周波数軸変換係数を平滑化バッファに登録する際に、発声される音韻による声道長の変動に伴う最適な周波数軸変換係数のばらつきを吸収するため、さらには推定精度の低下を防ぐため、今回の発声に対して推定された最適な周波数軸変換係数を評価し、平滑バッファに登録するか否かを判断する。その詳細について、図４を用いて説明する。

【００６３】前記の認識結果候補を求めた際用いた周波数軸変換係数α_0つまり前回までの発声に対して推定された最適な周波数軸変換係数α_0と、 今回の発声に対して推定された最適な周波数軸変換係数α_nを比較するにあたって、たとえば、 $|\alpha_0-\alpha_n|<0.100$ という評価関数（Ｓ１１０、Ｓ１１１）を用いて、この条件を満す場合にのみ今回の発声に対して推定された周波数軸変換係数α_nを採用・登録し、平滑化バッファ内の周波数軸変換係数を平均化し、周波数軸変換係数α_0を更新する（Ｓ１１２）。この周波数軸変換係数α_0は、次回の発声に対して、図2のＳ３０などでの周波数軸変換に用いられることになる。

【００６４】このように評価関数を用いることにより、推定された周波数軸変換係数のばらつきを抑える事ができることから、平滑化された周波数軸変換係数を精度よく求めることができるため、周波数軸変換を用いた話者正規化による効果をより高めることが可能となる。

【００６５】（実施の形態２）上記実施の形態１の音声認識装置は、一人の話者が発声していることを前提としているため、現在の発声話者に対して最適な周波数軸変換係数が推定され、この変換係数を用いて次回の発声に対して話者正規化が実施される。一方、家庭など複数の話者が交代で利用することが想定される場でこの音声認識装置を用いると、交代前話者に対して推定された周波数軸変換係数を用いて、交代後の話者に対して話者正規化が行われることになる。この時、交代直後に推定される周波数変換係数は 交代後の話者にとって必ずしも最適な値とは限らず、認識率を低下させる原因にもなり兼ねない。

【００６６】そこで、本発明の実施の形態２では、このように複数の話者が交代するような状況においては、話者交代を検出することによって、交代前話者の推定周波数変換係数の影響を受けずに、交代後話者に対して最適な周波数軸変換係数を求めることを可能とするものである。

【００６７】本実施の形態の音声認識装置における話者交代を検知する実施形態について、図６のフローチャートを用いて説明する。本実施形態は、実施の形態１の図２で説明した実施例に加えて、話者交代を検知する手段を有することを特徴としているもので、実施の形態１と説明の重複を省くために異なる部分のみを説明する。

【００６８】図２同様、Ｓ６０にて今回の発声に対して推定された最適な周波数軸変換係数に対して、過去所定回数の発声に対して推定された最適な周波数軸変換係数から算出される現話者に対して推定された最適な周波数軸変換係数とを比較して、今回の周波数軸変換係数を評価することにより話者の交代を検知し、話者交代を検知した場合には、前回までの発声に対して推定された最適な周波数軸変換係数保持している平滑化バッファを初期化し、新しい話者の周波数軸変換係数の登録を行う（Ｓ８０）。最後の処理として、図２同様に、現話者に対して推定された最適な周波数軸変換係数を算出する（Ｓ７０）。

【００６９】ここで、Ｓ８０における話者交代検出話者の処理の詳細について、図５を用いて説明する。話者の交代の検出は、前回の発声に対して推定された周波数軸変換係数α_{n-1} と 今回の発声に対して推定された周波数軸変換係数α_nを比較して行い、たとえば $|\alpha_n-\alpha_{n-1}|<0.150$という評価関数（Ｓ１２０、Ｓ１２１）を用いる。この条件を満たさない場合には、話者の交代を検出したとみなし、平滑化バッファの初期化を行うことにより（Ｓ１２２）、新しい話者に対する現話者に対して推定された最適な周波数軸変換係数α_nを出力する。

【００７０】このようにして、話者が交代しても認識率を低下させず、しかも従来法のように指定された単語の発声を促すこともなく、話者正規化を進めることが可能となる。

【００７１】（実施の形態３）次に、本発明の実施の形態３の音声認識装置について説明する。

【００７２】実施の形態１または実施の形態２では、一段階のマッチング法を用いたが、実施の形態３では2段階の認識方式について説明する。

【００７３】本発明の実施の形態３における音声認識装置の動作フローチャートを図７に示すが、実施の形態１または実施の形態２と異なる部分についてのみ説明するものとする。

【００７４】1段目の予備マッチングを行い（Ｓ５０）、

出力される認識結果候補のうち少なくとも１つから表現される音素系列に対して最適な周波数軸変換係数を推定した後（Ｓ６０）、話者交代を考慮して（Ｓ８０）、現話者に対して推定された最適な周波数軸変換係数を算出する（Ｓ７０）。さらにＳ７０にて求めた最適な周波数軸変換係数を用いて周波数軸変換を施した（Ｓ３１）特徴量を用いて、2段目の精密マッチング（Ｓ５１）を行う実施形態も可能である。

【００７５】（実施例）以上、本実施例の構成を用いて、１００単語を発声した男女５０名の音声データの認識実験を行った。まず、オンライン「教師なし」話者正規化を実現するために、変換係数平滑化の効果を調べる実験を行った。この実験においては、あらかじめ話者正規化を行わない条件で男女５０名について認識率を算出して、その結果認識性能の悪い１０話者を対象とした。

【００７６】図８に示す実験結果より、話者正規化学習は、評価発声データを７単語以上ではその効果がほぼ飽和していることから、評価発声データ１０単語を単位とすれば、オンライン「教師なし」話者正規化には十分効果があることが分かる。

【００７７】次に、音声区間情報に同期して周波数軸変換を行う効果について男女５０名について調べる実験を行った。なお、変換係数平滑化のためのオンライン学習データ数は１０とした。その結果（表１）に示すように、話者正規化を行う前は９３．７６％であったのに対し、音声区間情報を用いずに話者正規化を一律に行った場合、９４．７８％、本実施例に基づく有声音区間のみに話者正規化を行った場合は、９５．４４％に認識率が改善され、誤り率もそれぞれ約16.0％、約26.9％、改善された。

【００７８】

【表１】

	50名平均	最低10名	最低話者
話者正規化なし	93.76%	86.9%	77%
話者正規化あり	94.76%	89.4%	83%
話者正規化あり[有声区間のみ正規化]	95.44%	91.2%	87%

【００７９】また、50名の中で認識率の悪い10名についても効果が認められ、最低話者についても77％から、83％、87％と大幅に改善された。このことから、音声区間情報に同期して周波数軸変換を行なうことによる効果が認められる。

【００８０】なお、本実施例においては、単語マッチング方法として端点フリーのＤＰマッチング法を用いたが、ＨＭＭ（隠れマルコフモデル）での実施も可能である。

【００８１】なお、本実施例においては、単語マッチング時の距離尺度として共分散行列を共通化したマハラノビス距離を用いたが、共分散行列を共通化しないマハラノビス距離や、ＨＭＭから構成される音素モデルからを用いて計算することもできる。

【００８２】また、本実施例においては、認識対象を単語としたが、これを連続発声認識する際に利用することも可能である。

【００８３】なお、本実施例においては、音響的特徴を表現する特徴量としてＬＰＣメルケプストラム係数を用いたが、ＬＰＣケプストラム係数、メルケプストラム係数、ケプストラム係数での実施も可能である。

【００８４】なお、本実施例においては、入力される音声は、8kHzでサンプリングされたデータを用いたが、他のサンプリング周波数についても実施可能である．なお、本実施例においては、マッチングの際に音声区間情報は用いなかったが、音声区間情報を用いて、音声の始端を制限したマッチングなどを行うこともできる．

【００８５】

【発明の効果】本発明によれば、声道長差に起因するスペクトルの伸縮の影響を除去するため、入力音声のスペクトルに対して周波数軸変換を行なうにあたり、過去所定回数の発声から推定した変換係数の平均値を用いて、推定された周波数軸変換係数のばらつきの影響を抑えることにより、オンライン「教師なし」条件で話者正規化を実現し、高性能な音声認識装置を提供できるという効果を得る。さらに、有声音／無声音などの音声区間情報を利用して、声道特性に無関係な区間に対しては周波数軸変換を行わず 有声音区間にのみ周波数軸変換を行うことにより、周波数軸変換係数の推定をより精度のよく行なうことができるため認識率の向上を図ることが可能である。

【図面の簡単な説明】

【図１】本発明の実施の形態における音声認識装置のブロック図

【図２】本発明の実施の形態１における音声認識装置の処理フローチャート

【図３】（ａ）最適変換係数推定処理のフローチャート
（ｂ）尤度計算処理フローチャート

【図４】変換係数平滑化処理のフローチャート

【図５】話者交代検出処理のフローチャート

【図６】本発明の実施の形態２における音声認識装置の話者交代検出処理のフローチャート

【図７】本発明の実施の形態３における音声認識装置の２段階認識処理のフローチャート

【図８】変換係数学習データバッファの大きさと認識率を示す図

【図９】従来例の音声認識装置の処理フローチャート

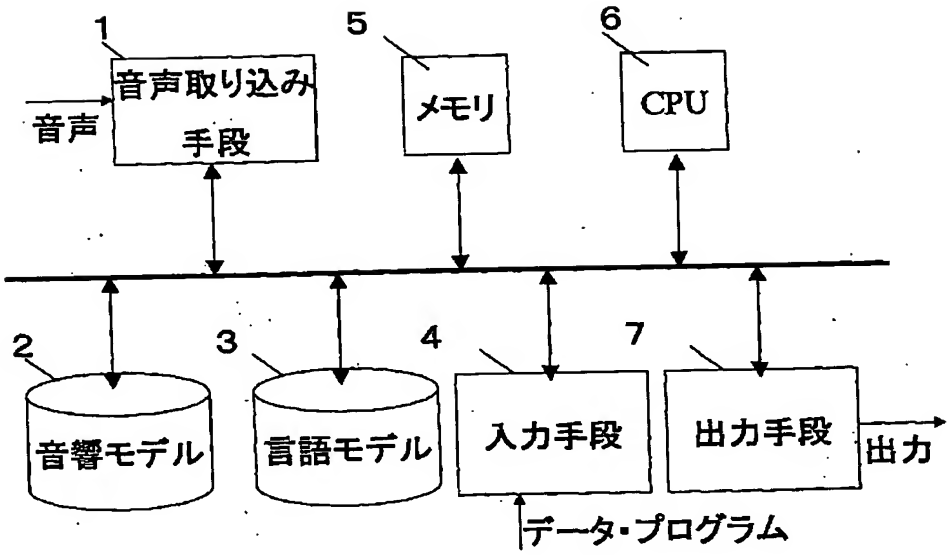
【符号の説明】

1　音声取り込み手段

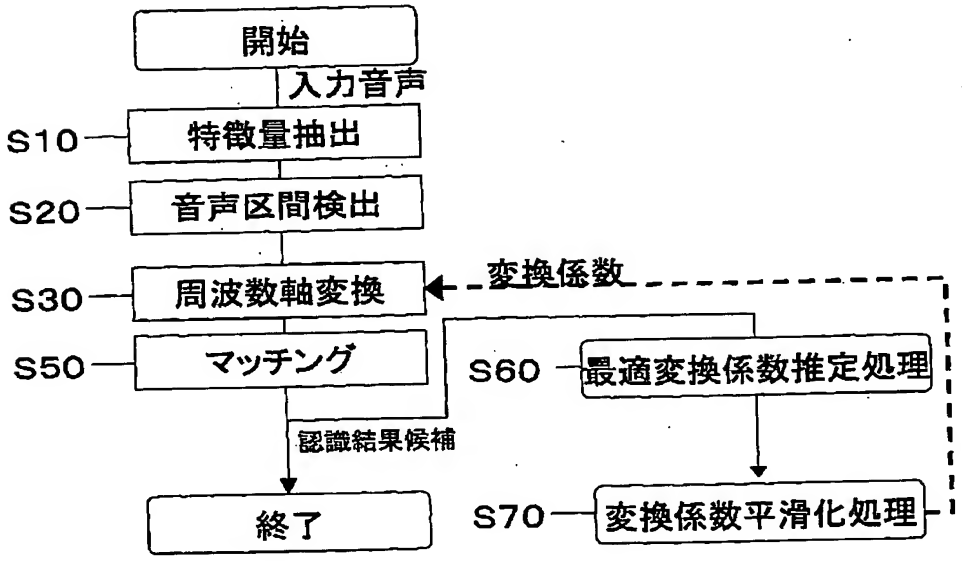
2　音響モデル

3　言語モデル
4　入力手段
5　メモリ
6　ＣＰＵ
7　出力手段

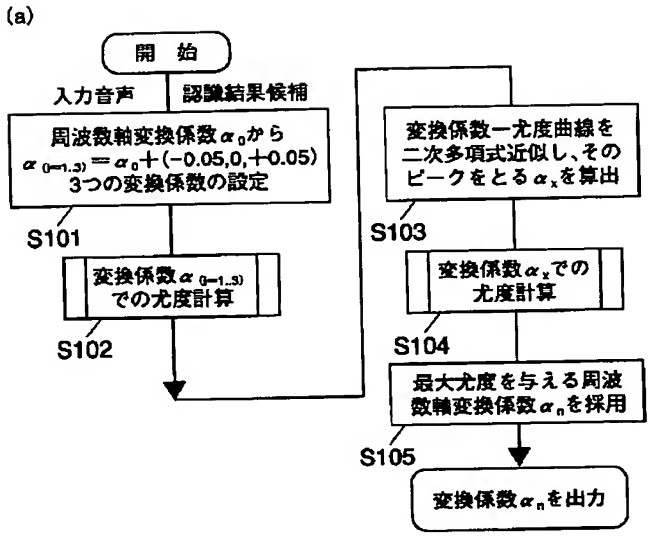
【図１】



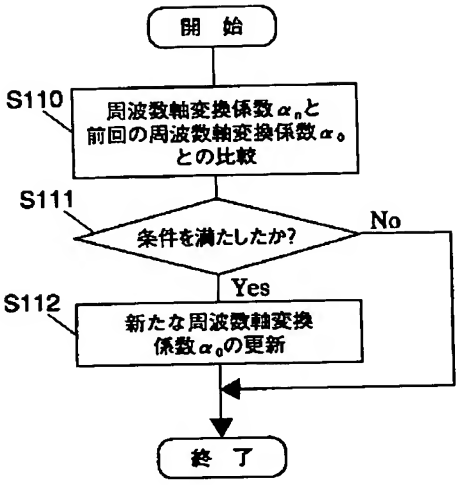
【図２】



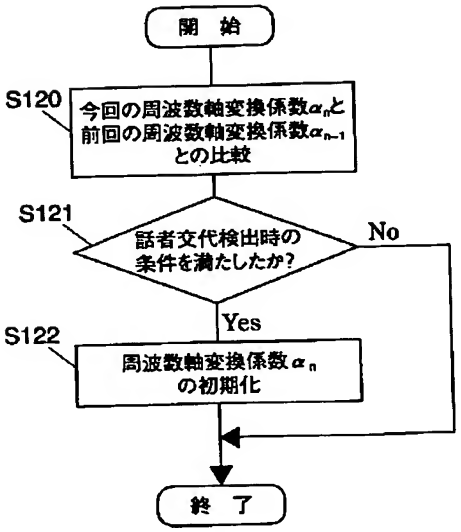
【図３】


(a)
開　始
入力音声
認識結果候補
周波数軸変換係数α_0から
$\alpha_{(i=1..3)}=\alpha_0+(-0.05,0,+0.05)$
3つの変換係数の設定
S101
変換係数$\alpha_{(i=1..3)}$
での尤度計算
S102
変換係数一尤度曲線を
二次多項式近似し、その
ピークをとるα_xを算出
S103
変換係数α_xでの
尤度計算
S104
最大尤度を与える周波
数軸変換係数α_nを採用
S105
変換係数α_nを出力
(b)
開　始
S106
周波数軸変換
S107
マッチング部
尤度出力

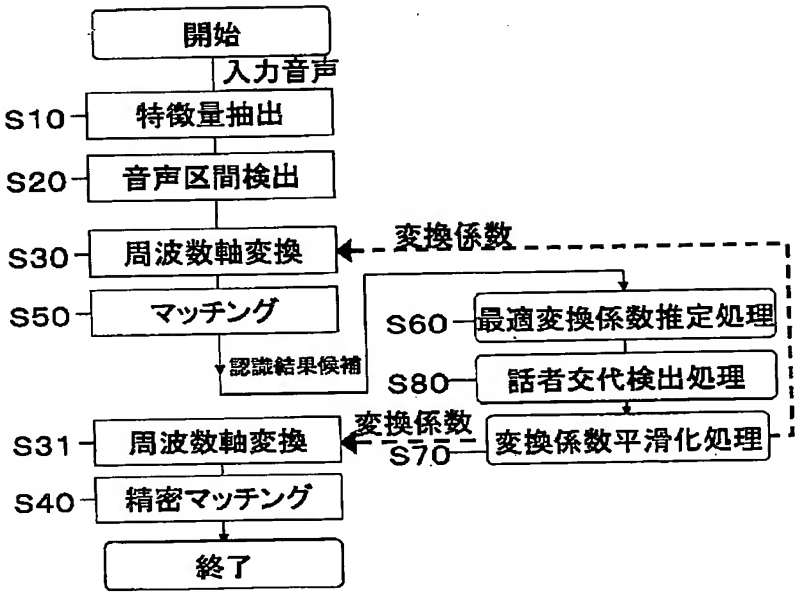
【図４】


開　始
S110
周波数軸変換係数α_nと
前回の周波数軸変換係数α_0
との比較
S111
条件を満たしたか?
No
Yes
S112
新たな周波数軸変換
係数α_0の更新
終　了

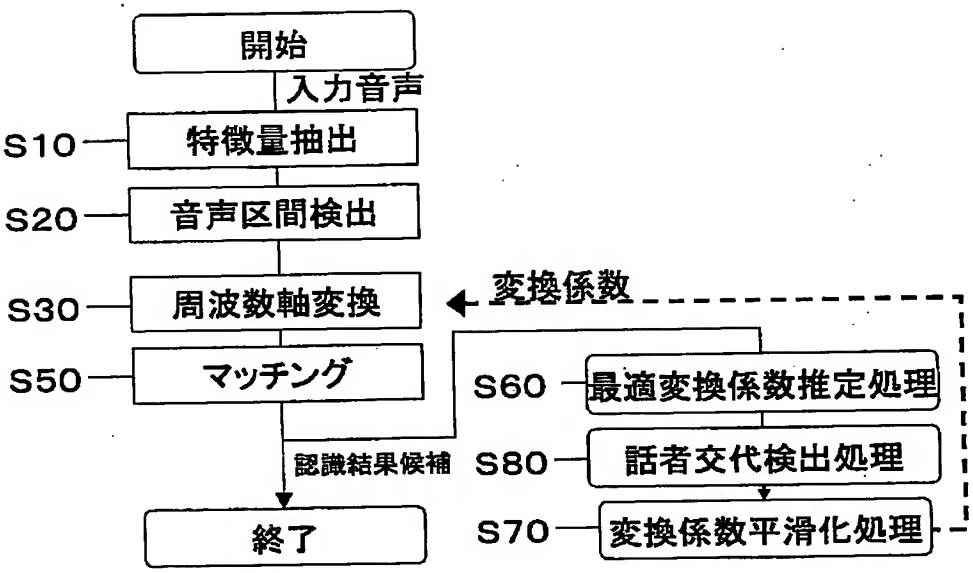
【図５】


開　始
S120
今回の周波数軸変換係数α_nと
前回の周波数軸変換係数α_{n-1}
との比較
S121
話者交代検出時の
条件を満たしたか?
No
Yes
S122
周波数軸変換係数α_n
の初期化
終　了

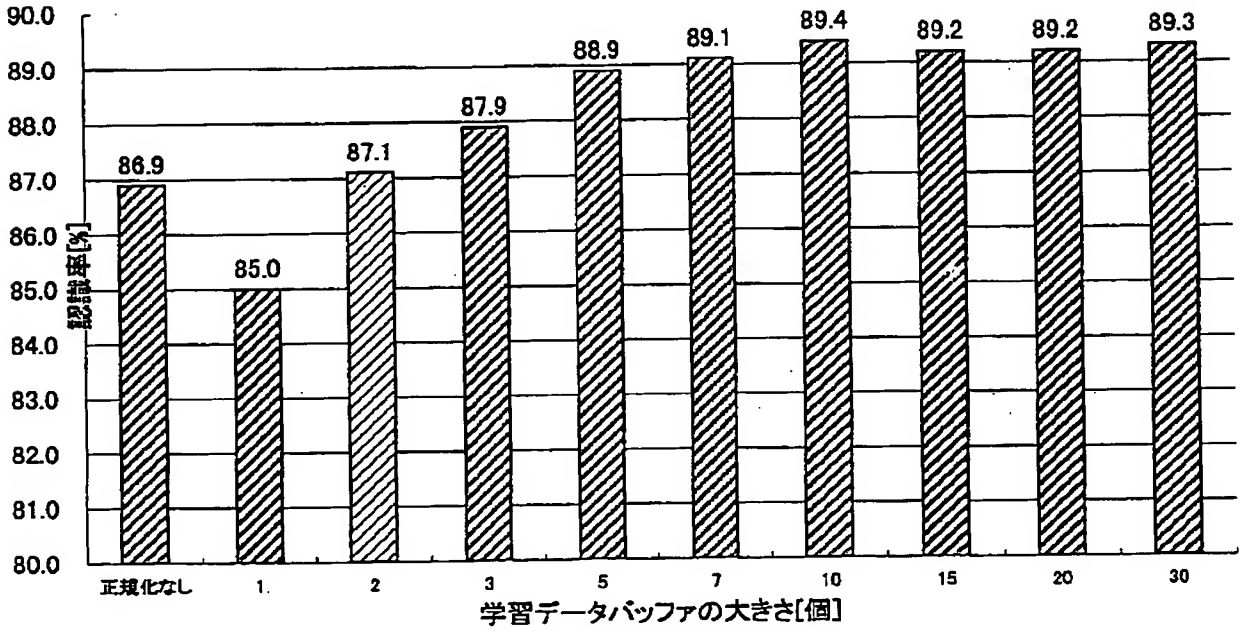
【図７】


開始
入力音声
S10
特徴量抽出
S20
音声区間検出
S30
周波数軸変換
変換係数
S50
マッチング
S60
最適変換係数推定処理
認識結果候補
S80
話者交代検出処理
変換係数
S31
周波数軸変換
S70
変換係数平滑化処理
S40
精密マッチング
終了

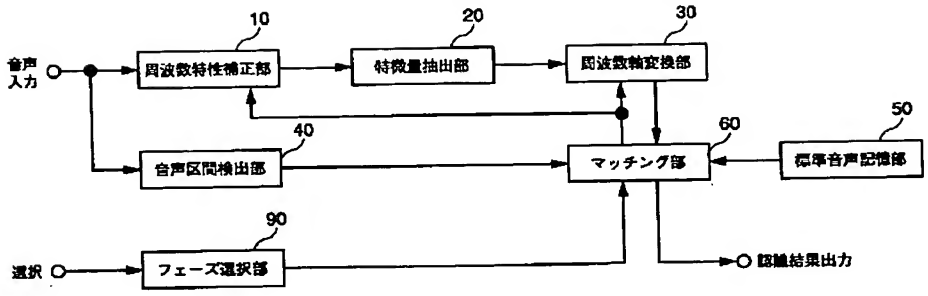
【図６】


開始
入力音声
S10
特徴量抽出
S20
音声区間検出
S30
周波数軸変換
変換係数
S50
マッチング
S60
最適変換係数推定処理
認識結果候補
S80
話者交代検出処理
終了
S70
変換係数平滑化処理

【図８】


90.0
89.0
88.0
87.0
86.0
85.0
84.0
83.0
82.0
81.0
80.0
認識率[%]
86.9
85.0
87.1
87.9
88.9
89.1
89.4
89.2
89.2
89.3
正規化なし
1
2
3
5
7
10
15
20
30
学習データバッファの大きさ[個]

【図９】



フロントページの続き

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ(参考)
// Ｇ１０Ｌ 101:16			

(72)発明者　木村　達也
神奈川県川崎市多摩区東三田３丁目10番１号　松下技研株式会社内

Ｆターム(参考)　5D015 AA02 BB02 FF07
9A001 BB06 EE05 GG01 HH16 HH17